Community Resilience Estimate
June 17, 2020

Community resilience is the capacity of individuals and households within a community to absorb, endure, and recover from the impacts of a disaster. The Community Resilience Estimates are experimental estimates produced using information on individuals and households from the 2018 American Community Survey (ACS) and the Census Bureau's Population Estimates Program as well as publicly available health condition rates from the National Health Interview Survey (NHIS).

The ACS is a nationally representative survey with data on the characteristics of the U.S. population. The sample is selected from all counties and county-equivalents in the U.S. and has a sample size of about 3.5 million addresses for the 2018 year survey. Publicly available data from the 2018 National Health Interview Survey (NHIS) are incorporated. The NHIS is the principal source of information on the health of the non-institutionalized population of the U.S. The NHIS is a cross-sectional household interview survey with about 35,000 households containing about 87,500 persons. Auxiliary data are used from the Population Estimates Program by tract, age group, race and ethnicity, and sex.

Modeling techniques used to develop the estimates are flexible and can be modified for a broad range of disasters (hurricanes, tornadoes, floods, etc.). These experimental estimates, in their current form, are specific to the current pandemic but could be modified to fit other disease outbreaks or weather–related disasters with differing risk factors. Local planners, policy makers, public health officials, and community stakeholders can use the estimates as one tool to help assess the potential resiliency of communities and plan mitigation strategies.

Resilience to a disaster is partly determined by the vulnerabilities within a community. In order to measure these vulnerabilities, we designed an individual risk index. Within this risk index, binary risk components are defined, adding up to 11 possible risks. A risk index is constructed using the weighted aggregate of risk factor. The specific measures we use are below.

<u>ACS-defined Risk Factors (RF) for Households (HH) and Individuals (I)</u>

- RF 1: Income-to-Poverty Ratio (IPR) < 130 (HH).
- RF 2: Single or zero caregiver household –only one or no individuals living in the household who are 18-64 (HH).
- RF 3a: Unit-level crowding - persons per room over 0.75 (HH)
- RF 4: Communications barrier –linguistically isolated or no one in the household with a high school diploma (HH)
- RF 5: No employed persons (HH)
- RF 7: Disability posing constraint to significant life activity (I)
- RF 8: No health insurance coverage (I)

<u>Domain defined RF</u>

- RF 6: Age >= 65 (I)
- RF 3: RF 3a = 1 or more persons reside within high-density tract (HH)

Health Condition RF

- RF 9: Serious heart condition (I)
- RF 10: Diabetes (I)
- RF 11: Emphysema or current asthma (I)

For each individual in the ACS microdata, the incidence rate from NHIS tables by health condition is used to randomly assign the risk factor (as a probability). The incidence rate for health conditions is estimated for 120 possible combinations (3 age groups by 2 sex groups by 5 race and ethnicity groups by 4 regions).

Weighted area-level tabulations of tallies and rates for each group are calculated, and accompanying direct replicate weight-based sampling variances are processed. Direct estimates and accompanying standard errors are produced and fed directly into a synthetic auxiliary proxy index. This index is used to assign indirect shares estimates, producing aggregate-level standard error calculations and subsequent shrinkage small area estimates.

The ACS-defined components are constructed and modeled jointly. Then, domain-defined variables are appended as a function of domain affiliation. Probabilities for the health risk components are derived from published NHIS tables by region, age group, sex, race, and Hispanic origin. Correlation among components are only to the level of their aggregation.

For each individual a set k of possible risk factors are defined. Each risk factor is binary, so let $\delta_{ik} \in \{1, 0\}$ be the outcome for risk factor k and individual i, with probability, P_{ik}. The outcomes are dependent. For example, consider the two risk factors of low income and unemployed. The probability of being low income is much higher if one is unemployed (and vice versa). At the micro level, we have a set of dependent Bernoulli-distributed variables at the micro-level.

We use micro level data to produce a measure at the domain level. This measure is the number of people facing a specific number of risk factors. For each domain, we produce k such measures.

Conceptually, at the person level, the measure is a sum of risk factors. For individual risk factor tallies of d = 0..k risk factors:

$$Y_{id} = I\left(\left(\sum_{k=1}^{K} \delta_{ik}\right) = d\right)$$

Where $= I(.)$ represents the indicator function, equals one if condition is true, zero otherwise. So for each individual, the outcome Y is vector of length k. Each element is a 1 or 0 representing whether that individual has that number of risk factors. These indicators represent a realization from a series of dependent Bernoulli distributions. This is generally termed a multivariate Bernoulli distribution.

Domain level

The domain-level outcomes we want to measure are, for domain m and individual risk factor sums of d = 0..k risk factors:

$$Y_{m,d} = \sum_{i=1}^{N_m} \{Y_{id}\}$$

So the quantity within the curly braces is a multi-variate Bernoulli. At the domain-level, $Y_{m,d}$ is multinomial.

Modeling Strategy

m = tract
j = age group x race-ethnicity group
$y_{mjk} = \sum_{i=1}^{n_{mj}} \delta_{ijk} w_{ijk}$ = ACS weighted estimate of positive flags for risk component k within domain j,
$ACSPOP_{mj} = \sum_{i=1}^{n_{mj}} w_{ijk}$ = ACS weighted estimates of tract-level pop within domain j,
POP_{mj} = auxiliary tract-level pop estimates within domain j.

Define 2 nested high-level estimation layers (b-layers within a-layers) for parameter estimation and calibration. Nationwide, a-layer is defined as 9 divisions crossed with 4 urbanization strata, and the b-layers are 4-15 subsets within each a, mostly defined along CBSA boundaries.

Each a-layer is large enough so that parameter estimates, and resulting synthetic estimates, are negligibly correlated with tract-level ACS estimates.

<u>Four step synthetic procedure</u>
Marginal post stratification estimates

For each a-layer-domain-risk component, edge rates are calculated.

$$r_{ajk} = \left(\sum_{m=1}^{M_a} y_{mjk}\right) / \left(\sum_{m=1}^{M_a} ACSPOP_{mj}\right)$$

These rates are applied to the tract-level population estimates by domain.

$$\hat{Z}^0_{mjk} = r_{ajk} * POP_{mj}$$

<u>Calibration</u>
An adjustment is applied to the b-layer estimates that result from aggregating the post-stratified tract-level estimates across tracts and demographic groups.

B-layer aggregates of the synthetic estimate across all demographic groups, $\hat{Z}^0_{bk}$ and resulting aggregate rates, and the corresponding direct estimate, r_{bk} are constructed.

$$\hat{R}^0_{bk} = \left(\sum_{m=1}^{M_b}\sum_{j=1}^{J}\hat{Z}^0_{mjk}\right) / \left(\sum_{m=1}^{M_b}\sum_{j=1}^{J} POP_{mj}\right)$$

The adjustment is calculated separately for each of the 36 a--layers, as un-weighted regression across the nested b-layers. Without the addition of further auxiliary data, this adjustment reduces to an intercept term. In other words it represents the difference in average rates between the initial post-stratified estimate and the ACS estimate.

$$\hat{\beta}_{0.a} = \frac{1}{B_a}\sum_{b=1}^{B_a} r^0_{bk} - \frac{1}{B_a}\sum_{b=1}^{B_a}\hat{R}^0_{bk}$$

$$\hat{R}^1_{bk} = \hat{\beta}_{0,a} + \hat{R}^0_{bk}$$

This mean adjustment can be larger than some of the 20 demographic-level post-stratification rates, r_{ajk}. Rather than apply $\hat{\beta}_{0.a}$ at the tract-level, a raking factor is calculated for each b-layer. This raking is then applied to the initial tract-level count.

$$\hat{Z}^1_{mjk} = \left(\hat{R}^1_{bk}/\hat{R}^0_{bk}\right)\hat{Z}^0_{mjk}$$

Regressions were designed at this high level of aggregation to avoid the prevalence of zeros at the tract-level, and maintain negligible correlation between synthetic and direct estimates at the tract level.

Outcome permutation stage

For a given risk component k^*, ACS weighted aggregates are used at the a level to calculate the post-stratification ratios for the 64 permutations (both $k^* = 0$ and $k^* = 1$ outcomes) of all other risk components. For. The result is two 64 by 1 empirical probability vectors, conditional on the k^* value, $r_{aj,k7}(k^* = 0)$ and $r_{aj,k7}(k^* = 1)$. Apply these ratio vectors to the counts obtained in part 2.b. above.

$$\hat{Z}^1_{mj,k7}(k^* = 0) = r_{a1,j,k7}(k^* = 0) * \hat{Z}^1_{mj}(k^* = 0)$$
$$\hat{Z}^1_{mj,k7}(k^* = 1) = r_{a1,j,k7}(k^* = 1) * \hat{Z}^1_{mj}(k^* = 1)$$

This results in a database of 128 by 20 age by rh groups by tract. This is repeated for all 7 of the dependent risk components. The final indirect estimate is the average of the 7 conditional estimates for each 128 by 20 age by rh groups by tract, thus $\hat{Z}^2_{mj,k7}$.

Domain-deterministic risks and target domain concept

The target concept is an aggregate of individuals by the number of risk factors, categorized into 3 groups: zero flagged risk factors, one to two flagged risk factors, and three or more risk factors.

For the health risk factors, the proportions for each domain (region by age group by sex by rh group) is obtained from published NHIS tables.

The first step is to reduce the database size by summing across risk factor categories. For the target concept we are only considering the number of risk factors faced.

$$\hat{Y}_{mjd} = I\left(\left(\sum_{k=1}^{K} \hat{Z}_{mjk}\right) = d\right), d = 0, \ldots, K$$

We split the resulting database by sex, as there are substantial differences in health risks by sex. The split uses the estimated population ratio of the sex in each age-race-ethnicity group, effectively setting the proportions for each of the prior risk factor groups equal between the sexes.

The proportions derived from NHIS tables are then applied to each domain, boosting the counts for higher risk numbers in each domain. This is repeated sequentially for each health condition.

The database size is further reduced by categorizing the RF-number by the 3 groups defined above.

$$\hat{Y}_{mj,dg}\ ;\ dg \in \{0RF, 1-2RF, 3PLRF\}$$

We aggregate by tract forming these counts and rates.

$$\hat{Y}_{m,dg} = \sum_{j=1}^{J} \{\hat{Y}_{mj,dg}\}$$

$$\hat{R}_{m,dg} = \hat{Y}_{m,dg}/POP_m$$

Synthetic MSE estimate

For the synthetic uncertainty estimate, we use a mean squared error approach relative to the average ACS sampling variance.

For the basic unweighted MoM estimator, and arbitrary collection of tracts, h, the equation is:

$$SynMSE_{h,dg} = \frac{1}{M_h}\left(\sum_{m=1}^{M_h} \left(\hat{R}_{m,dg} - r_{m,dg}\right)^2\right) - \frac{1}{M_h}\sum_{m=1}^{M_h} r\_gvfvar_{m,dg}$$

We use the following parameterization strategy. Assume the MSE for a given RF-group proportion estimate, $\hat{R}_{m,dg}$, is proportional to $\hat{R}_{m,dg}\left(1-\hat{R}_{m,dg}\right)$. And that the constant of proportionality can be assumed stable over a wide range of RF-groups and tracts. Under the assumption of stability of the multiplicative constant, calculate this constant at an aggregate level, averaging across both RF-groups and tracts.

Under this strategy, modify the previous MoM equation to average across RF-groups as well as tracts:

$$\mathrm{SynMSE_h} = \frac{1}{3\mathrm{M_h}} \sum_{\mathrm{dg}=0,1,3} \left(\sum_{\mathrm{m}=1}^{\mathrm{M_h}} \left(\widehat{\mathrm{R}}_{\mathrm{m,dg}} - \mathrm{r}_{\mathrm{m,dg}} \right)^2 \right) - \frac{1}{3\mathrm{M_h}} \sum_{\mathrm{dg}=0,1,3} \left(\sum_{\mathrm{m}=1}^{\mathrm{M_h}} \mathrm{r\_gvfvar}_{\mathrm{m,dg}} \right)$$

And calculate the estimated constant at this level of aggregation in two steps:

$$\mathrm{PmP_h} = \frac{1}{3\mathrm{M_h}} \sum_{\mathrm{dg}=0,1,3} \left(\sum_{\mathrm{m}=1}^{\mathrm{M_h}} \widehat{\mathrm{R}}_{\mathrm{m,dg}} \left(1 - \widehat{\mathrm{R}}_{\mathrm{m,dg}} \right) \right)$$

$$\mathrm{F_h} = \mathrm{SynMSE_h} / \mathrm{PmP_h}$$

Then for an individual tract:

$$\mathrm{SynMSE}_{\mathrm{m,dg}} = \mathrm{F}_{\mathrm{h:m} \in \mathrm{h}} \widehat{\mathrm{R}}_{\mathrm{m,dg}} \left(1 - \widehat{\mathrm{R}}_{\mathrm{m,dg}} \right)$$

This parameterization provides smooth, and stable, estimates that also satisfy the implicit constraints of the multinomial structure.

Composite Estimator

For two predictors of the same concept, a linear combination of the two, with the usual relative MSE weighting shown below, has approximate optimality if the covariance between the two is low relative to the individual MSE. This modeling strategy was designed to maintain low correlation.

$$\mathrm{w}_{\mathrm{m,dg}} = \mathrm{r\_gvfvar}_{\mathrm{m,dg}} / \left(\mathrm{r\_gvfvar}_{\mathrm{m,dg}} + \mathrm{SynMSE}_{\mathrm{m,dg}} \right)$$

$$\widetilde{\mathrm{R}}_{\mathrm{m,dg}} = \mathrm{w}_{\mathrm{m,dg}} \widehat{\mathrm{R}}_{\mathrm{m,dg}} + \left(1 - \mathrm{w}_{\mathrm{m,dg}} \right) \mathrm{r}_{\mathrm{m,dg}}$$

$$\mathrm{SE}\left(\widetilde{\mathrm{R}}_{\mathrm{m,dg}} \right) = \mathrm{sqrt}\left(\mathrm{w}_{\mathrm{m,dg}} \mathrm{SynMSE}_{\mathrm{m,dg}} \right)$$

Generalized Variance Function (GVF) of ACS estimates

For the synthetic variance estimation, $\mathrm{SynV_h}$, and subsequent composite estimator, we need a reliable sampling variance estimate. The GVF value was used for all shrinkage.

The basic GVF formula was:

$$\mathrm{r\_gvf}_{\mathrm{m,dg}} = \mathrm{F_m} \widehat{\mathrm{R}}_{\mathrm{m,dg}} \left(1 - \widehat{\mathrm{R}}_{\mathrm{m,dg}} \right)$$

Given un-weighted counts of ACS respondents within tract m, notated $\mathrm{n_m}$.
$\mathrm{F_m}$ is estimated with the regression,

$$\log\left(\text{r\_dirvar}_{m,dg}\right) - \log\left(\widehat{R}_{m,dg}\left(1 - \widehat{R}_{m,dg}\right)\right) = \alpha + \beta\log(n_m)$$

The regression was estimated for all m: $n_m > 25$.